Is firm growth random? A machine learning perspective

van Witteloostuijn, Arjen; Kolkman, Daan

published in
Journal of Business Venturing Insights
2019

DOI (link to publisher)
10.1016/j.jbvi.2018.e00107

document version
Publisher's PDF, also known as Version of record

Journal of Business Venturing Insights 10 (2018) e00107

Contents lists available at ScienceDirect

Journal of Business Venturing Insights

journal homepage: www.elsevier.com/locate/jbvi

Is firm growth random? A machine learning perspective

Arjen van Witteloostuijn[a,*], Daan Kolkman[b]

[a] VU Amsterdam, the Netherlands & University of Antwerp / Antwerp Management School, Belgium
[b] Jheronimus Academy of Data Science, the Netherlands

A B S T R A C T

This study contributes to the firm growth debate by applying machine learning. We compare a prominent machine learning technique – random forest analysis (RFA) – to traditional regression in terms of their goodness-of-fit on a dataset of 168,055 firms from Belgium and the Netherlands. For each of these firms, we have one to six years of historical data involving demographic and financial information. The data show high variation in firm growth rates, which is difficult to capture with traditional linear regression (R^2 in the range of 0.05–0.06). The RFA fares three to four times better, achieving a much higher goodness-of-fit (R^2 of 0.16–0.23). RFA indicates that perhaps firm growth is less random than suggested by traditional regression analysis. Generally, given the modest selection of variables in our dataset, this demonstrates that machine learning can be of value to firm growth research.

1. Introduction

After close to a century of research (Gibrat, 1931), firm growth is still a puzzle, by and large (Wennberg et al., 2016). The evidence seems to suggest that, at least in part, firm growth may be close to a random walk (Coad et al., 2013), may be the result of mere luck (Storey, 2011), or may feature a chaotic potpourri of different paths (Garnsey et al., 2006). Many theoretical frameworks have been put forward to predict firm growth rates. Similarly, a wide variety of empirical methods have been employed to tackle the problem, ranging from small-*n* case studies to large-*n* panel designs. But all to little or no avail: A firm growth model with a meagre R^2 of 0.15 is already a (relative) top performer, suggesting a high degree of randomness (Coad, 2009; Parker et al., 2010). In a recent *JBVI* exchange, Derbyshire and Garnsey (2014, 2015), and Coad et al. (2015) reflected upon the illusion of randomness versus Gambler's Ruin Theory debate. The former argue that the latter's conclusion of growth randomness is an artefact of their method, offering complexity science as an alternative lens.

The low goodness-of-fit of existing firm growth models stands in sharp contrast with that of state-of-the-art machine learning that can be used to, for instance, classify images or transfer speech to text. We build on the so-called “Data Science revolution” (Chen et al., 2012; McAfee and Brynjolfsson, 2012) to improve firm growth models’ goodness-of-fit, and to reflect upon the (illusion of) randomness debate. Specifically, we compare a prominent machine learning technique – random forest analysis – to two traditional parametric regression models (standard OLS and forward stepwise regression) in terms of their goodness-of-fit on a dataset of 168,055 firms from Belgium and the Netherlands. For each of these firms, we have one to six years of historical data with basic demographic and financial information.

* Corresponding author.
E-mail address: a.van.witteloostuijn@vu.nl (A. van Witteloostuijn).

https://doi.org/10.1016/j.jbvi.2018.e00107
Received 8 October 2018; Received in revised form 30 November 2018; Accepted 6 December 2018

2. Machine learning

As the volume, velocity, veracity, and variety of the data that society collects do increase rapidly, the analysis of such so-called Big Data transcends the cognitive capability of people (Kitchin, 2014). Consequentially, there is a considerable and growing reliance on algorithms to structure, analyze, and model data. The term "machine learning" broadly refers to algorithms that optimize model performance criteria – such as the traditional and well-established R^2 – by evaluating generated (or "predicted") output against observed (or "true") data. Machine learning, often used interchangeably with "artificial intelligence", is particularly helpful in cases where we do not have the knowledge required to formulate the rules of a target system or where such knowledge is tacit and cannot be readily transferred (Smola and Vishwanathan, 2014). Methods such as regression trees, neural networks, and support vectors have been around for quite for some time (Bishop, 2016). Machine learning moves beyond mere description. Typically, machine learning algorithms have the capacity to uncover non-obvious patterns in data, and facilitate reliable and accurate explanations and / or predictions.

We selected three models on the basis of their predominance in the entrepreneurship literature or their track record in the field of machine learning: ordinary least squares (OLS), forward stepwise regression (FSR), and random forest analysis (RFA). OLS is the standard parametric multivariate regression model that is very well known in academia, being its major empirical workhorse. FSR is a well-established procedure for the selection of independent variables in OLS regression. It proceeds by adding independent variables to the model one at a time. At each step, the independent variables not in the model are tested for inclusion in the model. The most significant of these variables is then added to the model, as long as its p-value is below some pre-set level. This threshold is typically set at 0.05. In addition, for this model, we removed independent variables with an in-time Pearson's non-autocorrelation higher than 0.8 or lower than −0.8, retaining the first of the two variables with a correlation exceeding this threshold.

RFA is a class of machine learning algorithms that can be used for regression and classification. An RFA involves a combination of multiple decision trees that are trained on different sub-sets of the data (Breiman, 2001). This approach of combining different models to improve performance is also referred to as "ensembles". In addition, using different sub-sets of the data, RFA determines the splits of the constituent decision trees by considering a random sub-set of predictor variables. Final predictions are acquired by aggregating across the constituent decision trees. This prevents the model from overfitting the data. RFA can detect non-linear and high-order interactions between determinants. Given space limitations, we cannot but briefly introduce the key intuition behind modern RFA. For insightful introductions, we refer to Boulesteix et al. (2012) and Loh (2011).

3. Data and methods

The Belgian-Dutch business data provider Graydon provided information that was collected as part of their regular business operations. So, in this paper, we have to work pragmatically with the available data, collected for different purposes than conducting academic research. However, to illustrate the potential added value of RFA, this Big Data set is appropriate. We return to this issue in the Discussion. We operationalized growth rate as an index of total assets growth:

$$Growth_i = \left(\frac{A_{it}}{A_{it-1}}\right)\cdot 100$$

Here A_{it} is the total assets in € of firm i at time t. For records with missing A values, the growth rate was computed based on the number of employees. This is a clear limitation of our data. Preferably, we would like to have been able to run separate analyses for different measures of growth (such as assets, employees and sales), as we know from extant work that different growth yardsticks are associated with different underlying causal processes (and hence different growth pattern outcomes; see, e.g., Coad et al., 2013; Derbyshire and Garnsey, 2015). However, the Graydon dataset is only relatively complete regarding total assets growth (and not at all in terms of employees or, even worse, sales).

Our initial dataset contains 2,494,784 records that are each associated with one firm in a sample of 533,626 unique enterprises. Due to missing values, not all firms could be used in our analysis. Moreover, given our interest in small and medium-sized firms, we removed outliers with a cutoff of five times the standard deviation. Moreover, this creates the needed level-playing field between RFA and OLS, as the latter is typically more sensitive to outliers.[1] The resulting number of records that can be used is 451,432. The number of unique firms is 168,055, of which 128,990 are Belgian and 39,065 Dutch.

The firms in our sample were labeled according to their industry using the second NACE level (divisions). After sifting through the variables, we ended up with a set of 16 "core" potential predictors in the form of contemporaneous and lagged demographic and financial measures. We added a one-year time lag for balance sheet total (or total assets), total equity, working capital, current ratio, solvency ratio, and number of employees. The legal person and sector are categorical variables, which we transformed into a series of binary dummies. This resulted in a total of 113 predictor variables.

To evaluate and compare the performance of the three techniques, we use three different datasets. *Dataset (1)* is the base dataset that includes 113 variables. *Dataset (2)* has all base variables, plus quadratic transformations of those variables and first-order product terms. *Dataset (3)* involves variables selected according to FSR. This results in the following set of eleven base predictors for the regression models: balance sheet total (t), balance sheet total (t-1), total equity (t), total equity (t-1), working capital (t), working

[1] We reran all analyses with these outliers included (available upon request), but this does not affect the patterns of results presented here. As expected, the performance of the OLS model decreased somewhat, to an R^2 of 0.04 on Dataset (1), while the RFA's performance was unchanged.

capital (*t*-1), current ratio (*t*), current ratio (*t*-1), solvency ratio (*t*), solvency ratio (*t*-1), and number of employees (*t*). With these three datasets, we can evaluate whether differences in model performance originate from the variable selection procedure or can be explained by the more flexible specification strategy of machine learning, which can model higher-order interaction effects.

We made five specific estimation choices. First, we standardize all three datasets by dividing the values by the L2-norm for that variable, using "least absolute deviations". Although not strictly necessary for traditional regression, such normalization is recommended for training purposes (LeCun et al., 2013). To prevent the normalization from becoming a factor in our comparative analysis, we initially used the normalized data as input for all the models we estimated. Second, we randomly split all three datasets in a training (80% of the data) and a validation set (20%). Whenever there is a large set of possible relationships, one has to be careful not to use the resulting freedom to find meaningless patterns in the data. This problem of overfitting is a very general phenomenon, and occurs even when the target function is not at all random. It afflicts every kind of learning algorithm. A typical approach to identify overfitting is to evaluate a model against a set of data that was not used to train the algorithm. This unseen dataset is also known as a hold-out set, validation or test set, and is also used to get a sense of how well a model will generalize (Chicco, 2017).

Third, we determine RFA's hyper-parameters by *k*-fold cross-validation on the training set of Dataset (1): *k*-fold cross-validation is a bootstrapping technique that draws random samples from the training set with replacement (Refaeilzadeh et al., 2009). To ensure robustness of our results, we employed 30-fold cross-validation. We first set an input list of hyper-parameters, or RFA-specific tuning parameters. Next, we conduct a so-called random search to identify hyper-parameters with the highest R^2. We use random search as opposed to grid search (Bergstra and Bengio, 2012). We set the number of randomly chosen combinations of hyper-parameters to 100. With this setting, random search finds a solution within 5% of the optimal solution 99% of the time. Subsequently, we select hyper-parameters following the 1-SE rule to "choose the simplest model whose accuracy is comparable with the best model" (Krstajic et al., 2014, p. 11). For our RFA, we selected the model with the least trees and highest number of values per leaf. Fourth, we fit RFA with the hyper-parameters to the full training set of Dataset (1), (2) and (3), again using 30-fold cross-validation. Fifth, we produce predicted firm growth rates for each of the three models, and evaluate the goodness-of-fit of these models on the training set and the validation set for all three datasets. All analyses were conducted in Python (3.5.4) using the scikit-learn (0.19.1) and Keras (2.1.2) packages.

4. Three techniques compared

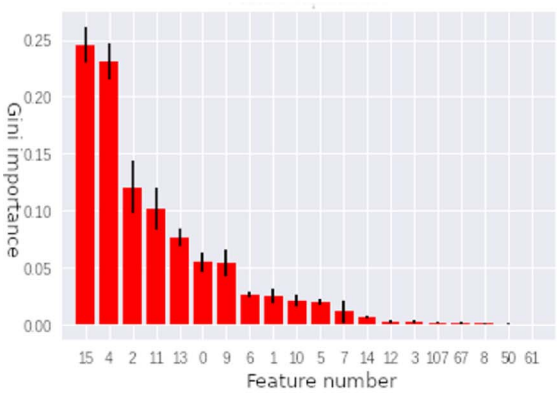
Following the recommendations of Legates and McCabe (1999), we consider the performance of our models on four criteria. The first is the coefficient of determination or R^2, which is the classic measure for how well the explanations or predictions approximate the observed values. The second is the Mean Absolute Error (MAE), which is the absolute difference between the explanations or predictions and the observed firm values. The third – Mean Squared Error (MSE) – is a similar statistic, measuring the average squared difference between the explained or predicted and actual or observed values. The fourth is the Root Squared Mean Error (RMSE), being the square root of the average of squared errors. In Table 1, we report the four fit statistics for our three models.

Clearly, the RFA performs best across all test statistics on Dataset (1). For instance, the R^2 of 0.23 (training set) or 0.16 (validation set) is impressive vis-à-vis the meagre R^2 of 0.05 or 0.06 for the traditional regression models. Importantly, the performance of both traditional regression models are almost identical (with an R^2 of 0.05 versus 0.06), implying that the much more flexible specification strategy employed by machine learning cannot explain the underperformance of our parametric multivariate regression benchmarks. The results on Dataset (2) confirm this. When we fit the models on Dataset (2), the performance of the RFA drops somewhat, but remains vastly superior. In Dataset (3), where we included first-order interaction effects and quadratic terms, the traditional methods start to catch up a little. This suggests that the superior performance of machine learning techniques originates from their flexible capacity to search for fit-enhancing higher-order interaction effects.

With an R^2 of 0.16 / 0.23, the RFA outperforms most top-fitting models in the traditional firm growth literature, even with our very limited set of basic demographic and financial explanatory variables – very impressive indeed. The loss in R^2 from the training to the test or validation set suggests slight overfitting of the data. Note that when fitted on a non-normalized dataset, the R^2 of the RFA on the training set decreased to 0.18, but the R^2 on the validation set increased to 0.17. In the following, we use the RFA fitted

Table 1
Model fit statistics.

	DATASET (1)				DATASET (2)				DATASET (3)			
TRAINING SET												
Model	R^2	**MAE**	**MSE**	**RSME**	R^2	**MAE**	**MSE**	**RSME**	R^2	**MAE**	**MSE**	**RSME**
OLS	0.06	17.25	781.63	28.22	0.05	17.63	829.10	28.80	0.08	17.28	79.54	28.22
FSR	0.05	17.56	821.69	28.27	0.05	17.56	821.69	28.27	0.06	17.48	821.05	28.66
RFA	0.23	15.76	665.61	25.80	0.14	16.78	749.38	27.37	0.17	16.45	722.47	26.88
VALIDATION SET												
Model	R^2	**MAE**	**MSE**	**RSME**	R^2	**MAE**	**MSE**	**RSME**	R^2	**MAE**	**MSE**	**RSME**
OLS	0.06	17.25	781.63	28.22	0.05	17.71	828/67	28.79	0.03	17.48	825.86	29.10
FSR	0.05	17.64	821.90	28.67	0.05	17.64	821.90	28.67	0.04	17.37	824.11	28.98
RFA	0.16	16.69	733.36	27.08	0.13	16.96	760.00	27.59	0.15	16.81	746.35	27.31



Ranking	Description	Feature #
1	Total Assets	15
2	Working Capital	4
3	Total Equity	2
4	Current Ratio	11
5	Solvency Ratio	13
6	Number of Employees	0
7	Firm Age	9
8	Active (Dummy)	6
9	Country (Dummy)	1
10	Total Assets (t -1)	10
11	Working Capital (t-1)	5
12	Total Equity (t-1)	7
13	Current Ratio (t-1)	14
14	Solvency Ratio (t-1)	12
15	Number of Employees (t-1)	3
16	Total Asset Growth (t-1)	107
17	NACE2 Code = 1 (Dummy)	67
18	NACE2 Code = 10 (Dummy)	8
19	NACE2 Code = 11 (Dummy)	50
20	NACE2 Code = 12 (Dummy)	61

Fig. 1. Plot of relative importance of firm growth's explanatory variables. All variables measured at *t*, except if indicated otherwise. Gini importance is defined as the total decrease in node impurity averaged over all trees in the random forest. A high Gini importance indicates the given feature has an important contribution to the predicted outcome.

on this non-normalized dataset to provide more details on the model's mechanics. We do so partly because the R^2 of this model in the validation set is higher, but also because normalized data can be hard to interpret.

The relative importance of the explanatory variables in the RFA can be measured using the mean decrease in accuracy, or the percentage increase in the MSE. This measure corresponds to the difference between the MSE for including and excluding that variable, averaged over all the trees and divided by the standard deviation of the differences. Machine learning's output gives so-called “feature importance”, which indicates the relative weight of each of the listed variables – or “features” in machine learning terminology – in explaining or predicting the “target variable”. The five most important variables for the RFA in descending order are: Total assets, working capital, total equity, current ratio, and solvency ratio, all in *t*. The complete list is provided in Fig. 1.

The RFA reveals that all the contemporaneous financial variables (as a group) are more important than all the demographic (ranked in-between) or lagged measures, the latter ranking – as a cluster –at the bottom. However, as the nature of Fig. 1's list makes clear, machine learning is not a parametric method. The standard RFA output does offer insight into the relative importance of all explanatory variables, and it does provide a statistic for the increase in fitness due to adding a specific explanatory variable, but this is different from the familiar β-coefficients and *p*-values produced in traditional parametric techniques in econometrics. So, the substantially higher overall fitness, an R^2 of ~ 0.05 for the traditional regression methods vis-à-vis ~ 0.17 for the RFA, is traded off against lack of insight in parametric effect sizes and significance values.

5. Discussion

This study examines whether the performance of firm growth models can be improved by applying modern machine learning. We constructed a large Big Data set with 168,055 unique firms, each associated with information for one to six years. Our analysis demonstrates that the random forest analysis (RFA) performs best on the training and validation set, and much better so than traditional multivariate regression, with an R^2 of 0.16 (the validation set) or 0.23 (the training set) for RFA vis-à-vis an R^2 of ~ 0.05 (both sets) for the traditional regression models. The goodness-of-fit of the RFA is on par with or superior to that of top contenders in the firm growth field, despite the very basic set of demographic and financial variables in our dataset. The extant models in the firm growth literature that perform in the R^2 ≈ 0.15 range include a much larger selection of non-demographic and non-financial information, typically adding data on personality and strategy measures collected through surveys.

Importantly, the RFA outperformed the multivariate OLS and forward stepwise regression models that were estimated in a traditional parametric fashion. The RFA retains its lead even when a smaller subset of variables is used. This shows that the performance gain of the RFA can be attributed to the machine learning technique, and cannot be solely explained by the larger flexibility of the machine learning algorithm. Only when interaction effects are added to the other models manually, do they start to catch up somewhat in terms of R^2. The superior performance of the RFA can thus be attributed to its flexible capacity to represent higher-order interact effects. Ultimately, this suggests that firm growth cannot be readily explained by a simple set of variables. Rather, firm growth exhibits subtle interaction effects and non-linearities, which are two key features of complex systems (cf. McKelvey, 2004). This supports Derbyshire and Garnsey's (2014, 2015) argument that, seen through a complexity science lens, firm growth is not random.

This immediately points to an intriguing issue regarding another key aspect of complexity theory: sensitivity to initial conditions (StIC). Neither RFA nor any traditional regression method is able to handle this well, except with infinitely accurate measures of these initial conditions. In entrepreneurship and firm growth studies, such ultimate measurement perfection is not within reach. However, given that our RFA of firm growth is associated with substantive explanatory power, even with a very limited set of potential predictors, does suggest that StIC may not be an issue here. If so, much of firm growth is not random after all, albeit driven

by complex and non-linear higher-order interactions among a large set of predictors. Of course, our study cannot, in any way, provide definite evidence. Future research is needed, replicating what we have done here, as well as extending this line of work by adding a richer set of predictors, as well as other measures of firm growth.

Finally, intriguingly, machine learning is not widespread in entrepreneurship studies. Ironically, this is in sharp contrast with the huge popularity of machine learning in the real business world. For instance, Google and Facebook would have been nowhere without machine learning, which is the very engine (with Big Data as the fuel) keeping their business model going (and impressively so). The key benefit of machine learning is high predictive accuracy, a feature vitally important in many corners of business life. This is not different in the domain of real-life entrepreneurship, in all cycles across an enterprise's lifecycle.[2] For instance, in the start-up phase, machine learning can be used to deeply analyze market potential, and during the scale-up phase, machine learning can guide lean experimentation to further develop the SME's product portfolio. Future research should (and probably will) open the black box of machine learning algorithms further to, hopefully, produce output that can match that of traditional parametric econometrics, generating statistics similar in nature to β-coefficients (economic significance) and p-values (statistical significance). An example is Papagionnopoulou et al. (2017), developing a non-linear Granger causality RFA framework. For now, we suggest to combine machine learning with traditional parametric techniques, as both sets of methods are associated with complementary strengths (and weaknesses).

Acknowledgements

The data used in this study were provided by Graydon the Netherlands and Graydon Belgium. Graydon is not responsible for any of our conclusions. We are grateful to Bas Bosma, Joeri De Caigny, Wim Coreynen, Pourya Darnihamedani, Marcus Dejardin, Barry Delhez, Julie Hermans, Maurits Kaptein, Simon Noyons, Wouter Stam, Eric Vandenbroele, Diemo Urbig, Johanna Vanderstraeten, Jack van Wijk, Mark Zwart and an anonymous reviewer for their comments and suggestions. Of course, all errors are ours.

Conflict of interest

None.

References

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13 (Feb), 281–305.
Bishop, C., 2016. Pattern Recognition and Machine Learning. Springer Press, New York.
Boulesteix, A.-L., Janitza, S., Kruppa, J., König, I.R., 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. WIREs Data Min. Knowl. Discov. 2, 493–507.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Chen, H., Chiang, R.H.L., Storey, V.C., 2012. Business intelligence and analytics: from big data to big impact. MIS Q. 36 (4), 1165–1188.
Coad, A., 2009. The Growth of Firms: A Survey of Theories and Empirical Evidence. Edward Elgar, Cheltenham.
Coad, A., Frankish, J., Roberts, R.G., Storey, D.J., 2013. Growth paths and survival chances: an application of Gambler's Ruin theory. J. Bus. Ventur. 28 (5), 615–632.
Coad, A., Frankish, J., Roberts, R.G., Storey, D.J., 2015. Are firm growth paths random? A reply to “firm growth and the illusion of randomness”. J. Bus. Ventur. Insights 3, 5–8.
Chicco, D., 2017. Ten quick tips for machine learning in computational biology. BioData Min. 10 (1), 35.
Derbyshire, J., Garnsey, E., 2014. Firm growth and the illusion of randomness. J. Bus. Ventur. Insights 1–2, 8–11.
Derbyshire, J., Garnsey, E., 2015. Are firm growth paths random? A further response regarding Gambler's Ruin Theory. J. Bus. Ventur. Insights 3, 9–11.
Garnsey, E., Stain, E., Hefferman, P, 2006. New firm growth: Exploring processes and paths. Ind. Innov. 13 (1), 1–20.
Gibrat, R., 1931. Les Inegalites Economique. Recueil Sirey, Paris.
Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. J. Cheminform. 6 (1), 1–10.
Kitchin, R., 2014. Big Data: new epistemologies and paradigm shifts. Big Data Soc. 1 (1), 1–12.
LeCun, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models. Proc. ICML 30 (1), 1–3.
Legates, D.R., McCabe, G.J., 1999. Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35 (1), 233–241.
Loh, W.-Y., 2011. Classification and regression trees. WIREs Data Min. Knowl. Discov. 1, 14–23.
McAfee, A., Brynjolfsson, E., 2012. Big Data: the management revolution. Harv. Bus. Rev., 60–68.
McKelvey, B., 2004. Toward a complexity science of entrepreneurship. J. Bus. Ventur. 19 (3), 313–341.
Papagionnopoulou, C., Miralles, D.G., Decubber, S., Demuzere, M., Verhoest, N.E.C., Dorigo, W.A., Waegenan, W., 2017. A non-linear granger-causality framework to investigate climate-vegetation dynamics. Geosci. Model Dev. 10, 1945–1960.
Parker, S.C., Storey, D.J., van Witteloostuijn, A., 2010. What happens to gazelles? The importance of dynamic management strategy. Small Bus. Econ. 35 (2), 203–226.
Refaeilzadeh, P., Tang, L., Liu, H., 2009. Cross-validationEncyclopedia of Database Systems. Springer, Chicago, IL, 532–538.
Smola, A., Vishwanathan, S., 2014. Introduction to Machine Learning. Cambridge University Press, Cambridge, UK.
Storey, D.J., 2011. Optimism and chance: the elephants in the entrepreneurship room. Int. Small Bus. J. 29 (2), 303–322.
Wennberg, K., Delmar, F., McKelvie, A., 2016. Variable risk preferences in new firm growth and survival. J. Bus. Ventur. 31 (4), 408–427.

[2] See, e.g., http://www.nielsen.com/us/en/insights/news/2014/how-small-businesses-can-scale-the-big-data-barrier.html.